Eesti vahekeele korpuse töörühma V sügisseminar “Esimese ja teise keele omandamine ning korpusuuringud”
27.10.2010.
Tallinna Ülikool

Vene õppijakeele korpuse koostamise põhimõtted

Alisa Shmeleva

Korpuse mõiste

- *Korpuse* mõiste on mitmetähenduslik ja ebaselgelt kasutatud.
- Üldkasutuses mõistetakse *korpuse* all mistahes tekstikogu, mis on elektroonsel kujul.
- Lingvistikas viimasel ajal *korpuse* all mõistetud mitte lihtsalt teksti (ingl *running text*), vaid keelematerjali kogu, mis on spetsiaalselt valitud ühel või teisel uurimistöö eesmärgil (Копотев 2003: 37 – 38).

Korpuse mõiste

- Kadri Muischneki järgi on „korpus polüfunktsionaalne elektroonsel kujul olev tekstikogu, millesse kuuluvad tekstid on valitud eesmärgipäraselt, nii et nendest koosnev tervik annaks tõepärase pildi kogu keelest“ (Kadri Muischneki loengud: (http://www.cl.ut.ee/kursused/korp_ling01, 22.10.2010).

Korpust iseloomustab

- kindlaspiiriline tekstivalik ja representatiivsus (ingl *sampling* & *representativeness*)
- piiratud või vaba suurus ehk maht (ingl *finite* and *non-finite size*)
- arvutis nähtav ja töödeldav keeleaines (ingl *machine-readable form*)
- standardmärgendusega tekstikogu (ingl *a standard reference*)

(McEnery & Wilson 1996: 29)

Korpuse mõiste

- **Niisiis on *korpus*** kirjalike tekstide või transkribeeritud kujul suulise kõne elektroonne kogu, mida kasutatakse keele uurimise ja kirjeldamise alusena (Kennedy 1998: 1).

Mis on õppijakeel ja õppijakeele korpus?

- Mare Kitsnik on *õppijakeelt* defineerinud kui keelevarianti, mille õppija ise loob õppimise ajal ning milles ta kaldub rohkemal või vähemal määral standardist kõrvale (Kitsnik 2007: http://evkk.tlu.ee/pdfs/Kitsnik_artikkel151106.pdf, 22.10.2010).

Mis on õppijakeel ja õppijakeele korpus

- *Õppijakeelt* on nimetatud ka *vahekeeleks* (ingl *interlanguage*), mille õppija loob oma emakeele (K1) ja õpitava keele (K2), mõnikord ka juba omandatud võõrkeel(t)e, emakeele ning uue õpitava keele alusel“
- *Vahekeelt* on nimetatud ka sihtkeele variandiks, mis mõningal määral kaldub normist kõrvale.

Mis on õppijakeel ja õppijakeele korpus

- Ka *õppijakeele* ehk *vahekeele korpus* (ingl *learner coprus* ja *interlanguage corpus*) on elektroonne keeleressurss, nagu ka *teise keele korpus* (ingl *L2 corpus*).
- Teisitisõnu: õppijakeele korpus on samasugune elektroonne keeleainese kogu nagu igasugune muu keelekorpus, ainult et õppijakeele korpuse algmaterjali autoriteks on keeleõppijad.

- Eesmärk on koostada vene õppijakeele korpus, mis sisaldab gümnaasiumiõpilaste kirjalikke tekste, milles vene keelt on kasutatud esimese ja teise keelena.
- sisaldab peamiselt gümnaasiumiõpilaste kirjandeid ja esseesid ning eksamikirjandeid.

Korpuse koostamise etapid

Graeme Kennedy järgi on korpuse koostamisel kolm peamist etappi:

1) korpuse kavandamine (ingl *corpus design*),

2) tekstide kogumine (ingl *text collection*) ja andmete salvestamine (ingl *capture*),

3) teksti koodering (ingl *text encoding*), annoteerimine (ingl *annotation*) ja märgendamine (ingl *text markup*)

(Kennedy 1998: 70).

Korpuse kavandamine (ingl *corpus design*)

- Korpuse kavandamine sõltub otseselt sellest, milleks korpus on mõeldud ja kuidas korpus oleks võimalik tulevikus kasutada.
- Tavaliselt kasutatakse korpust keelekasutuse analüüsimiseks ning kirjeldamiseks.

Korpuse kavandamine

- Mis tüüpi korpus on?
- Suletud (Nt Brown corpus, 1 mil sõna) või avatud (monitorkorpus)
- Sünkroonne (staatiline) või diakrooniline (dünaamiline)
- Standardkorpus või spetsiaalne

Korpuse liigid

- *Suletud korpus* on selline, mille tekstide pikkus on kindel suurus ja kust ei saa tekste välja jätta ega juurde lisada. Sellised korpused on representatiivsed. Esimese põlvkonna korpused oli sel viisil koostatud. Näiteks, Brown korpus, mis on piiratud suurusega (1 miljon sõna).

- *Avatud korpus* on selline, millest võib vajadusel osa tekste välja jätta, neid sinna juurde lisada või välja vahetada. Neid korpusi nimetatakse ka *monitorkorpusteks*. Nende korpuste struktuur ja suurus on pidevalt muutuv.

Korpuse liigid

- Kui korpus kajastab mingit ajahetke keele arengus, siis nimetatakse seda *sünkroonseks*
- Kui kajastab erinevate aegade keelekasutust ja selle ajalisi muutusi, siis peetakse silmas *diakroonilist korpust*

Korpuse liigid

- *Tasakaalustatud korpuseks* nimetatakse korpust, mis esindab võrdselt kõiki allkeeli.
- *Spetsiaalne korpus* keskendub ühele kindlale keelevariandile

Korpuse kavandamine
Vene õppijakeele korpus

- Vene õppijakeele korpus on kavandatud Eesti vahekeele korpuse (EVKK) allkorpusena.
- Korpus on spetsiaalne ja avatud ehk monitorkorpus, mis on koostatud kindlal uurimiseesmärgil:
- võrrelda vene keele kui esimese (K1) ja teise keele (K2) ühisjooni ja kasutuserinevusi Eestis.
- Kavas on võrrelda Eestis elavate ja gümnaasiumis õppivate noorte vene keele kui K1 ja K2 kasutamist.
- See on ssünkroonne korpus

Korpuse kavandamine

- Korpuse koostamine on seotud minu doktoritöö teemaga *Morfosüntaksiliste konstruktsioonide kasutus vene õppijakeeles*
- Analüüsin morfosüntaksilisi konstruktsioone, mida kasutavad aktiivselt vene emakeelega kõnelejad ja eesti emakeelega vene keele õppijad.

Korpuse kasutamise eesmärk doktoritöös

- Korpuse kasutamise peamiseks eesmärgiks on välja selgitada, mis on vene õppijakeelele tüüpiline (Sinclair 1995: 17) ja mille poolest see erineb või sarnaneb standard vene keelega ehk vene kirjakeelega.
- Selleks vajalik keeleaines leitakse võrdluskorpusest ehk referentskorpusest, millena on kavas kasutada Helsingi Ülikooli HANCO-korpust (The Helsinki Annotated Corpus of Russian Texts HANCO - Russian 22.10.2010), sest see on väga hoolikalt koostatud, representatiivne, morfoloogiliselt ja süntaktiliselt märgendatud ning kontrollitud.

  vt http://www.ling.helsinki.fi/projects/hanco/mte/mockyMSD.html ja http://www.helsinki.fi/venaja/english/e-material/hanco/index.htm (22.10.2010)

- Uurimus on korpuspõhine (ingl *corpus-based language analysis*);
- Korpus on analüüsitava keeleainese allikas
- Tekstimaterjali töödeldatakse erinevate arvutiprogrammide (nt Word SmithTools) ja keeletarkvara (nt vene keele morfoanalüsaator) abil.
- Saadud andmeid interpreteeritakse lingvistiliselt

Korpuse kavandamine

- Doktoritöö ning vene õppijakeele korpuse koostamine on seotud riikliku programmi „Eesti keel ja kultuurimälu (2009-2013)“ projektiga „REKKi käsikirjaliste materjalide digiteerimine, Eesti vahekeele korpuse alamkorpuste loomine ja korpuse kasutusvõimaluste populariseerimine (2009-2013)“.

Toetudes EVKK-le, võib vene õppijakeele allkorpust kasutada:

- igaüks, kel on huvi vene õppijakeele vastu
- tulevikus on see õppekeskkond vene keele õppijale, õpetajakoolituses ja täiendõppes osalejale
- üliõpilaste, magistrantide ja doktorantide töökeskkond (vene ja eesti õppijakeele (kõrvutava) uurimine)
- soovi korral võib korpuses luua individuaalse töökeskkonna ning arendada korpust vastavalt kitsamale uurimisteemale, lisades uusi märgendusi ja täpsustades üldist veaklassifikatsiooni
- kasutades korpuse materjali, on oluline korpusele viidata

Tekstide kogumine ja andmete salvestamine (ingl *text collection & capture*)

- Teine etapp korpuse koostamisel on tekstide kogumine ja andmete salvestamine
- Enne korpuse koostamist on vaja läbi mõelda,
  - missuguseid tekste koguda (teskti liik)
  - missuguseid andmeid koguda (nt teksti autori kohta)
  - kas kasutada juba avaldatud dokumente
    - Interneti ajastul on need saadaval elektroonsel kujul ega vaja digiteerimist

(vt Stefanowitsch: http://www-user.uni-bremen.de/~anatol/docs/corp_compilation.pdf, 29.04.2010)

Tekstide kogumine ja andmete salvestamine

- Kuna see meil on tegu vene õppijakeele korpuse koostamisega, siis on korpusse kavas koondada klassis ja kodus kirjutatud kirjandid ja eksamitööd.
- Kirjalikke tekste on kergem koguda ning valdav osa korpusi kajastab kirjalikku keelekasutust.

Tekstide kogumine ja andmete salvestamine

- Sihtgrupp on Eesti koolide gümnaasiumiõpilased.
- Siia kuuluvad nii vene õppekeega koolid, kuid ka eesti õppekeelega koolid, kus õpitakse vene keelt, või siis segatüüpi koolid nagu näiteks Vanalinna Riigikool Narvas.
- Korpuse tekstivalik on piiratud ka regionaalselt: Narva ja Tallinna koolid.

Tekstide kogumine ja andmete salvestamine

- Barlow, Barkhuizen ja Sõrmus on esile toonud, et keelekasutust mõjutavad teksti autori (õppija) vanus, sugu, haridustase, keelekeskkond, õppija emakeel ja pätiolu, sotsiaalne kuuluvus, teised õpitavad keeled ning nende osaoskuste tase, varemõpitud keeled ja keele prestiizsus

(vt Kadri Sõrmus 2008: 18 (https://dspace.utlib.ee/dspace/bitstream/10062/6217/1/sormus_kadri.pdf, 22.10.2010).

Tekstide kogumine ja andmete salvestamine

- Seega sisaldab vene õppijakeele allkorpus metateavet tekstid kirjutanud õpilaste kohta:
- vanus,
- klass,
- emakeel jne.
- EVKK veebilehel on olemas metainfo kogumiseks mõeldud ankeet, mille eestikeelse variandi tõlkisime vene keelde, täpsustades ning lisades küsimusi, näiteks küsimus keelekeskkonna suhtes:
- Mis keelt kasutatakse kodus, sõpradega suheldes, vaba aja veetmisel või teiste keelte valdamine jne
- Need täpsustavad küsimused on esitatud selleks, et välja selgitada, missugused välistegurid võivad mõjutada õppilaste keelekasutust ning veatekkimist

Ankeet õpilaste jaoks eesti keeles

Informatsioon õpilase kohta:

- Nimi:
- Vanus:
- Sugu:
- Emakeel:
- Kool:
- Haridustase (alg-, põhi, gümnaasium), klass:
- Õppekeel koolis:
- Teised õpitavad keeled: esimene, teine, muud:
- Keelekeskkond:
- Kodune keel:
- Sõbrad:
- Vaba aja veetmine (mängud, filmid, raamatud jne):
- Elukoht (maakond, linn, linnaosa, tänav):

Ankeet õpilaste jaoks vene keeles

- Информация об ученике:
- Имя:
- Возраст:
- Пол:
- Родной язык:
- Школа:
- Уровень образования (начальная школа, основная, гимназия), класс:
- Язык обучения в школе:
- Знание других языков, первый иностранный язык, второй:
- Социальное происхождение:
- Языковая среда
- Домашний язык:
- Друзья:
- Свободное время (игры, фильмы, книги и т.д.):
- Местожительство (уезд, город, часть города, улица):

Tekstide kogumine ja andmete salvestamine

- Andmete kogumisel ja tekstide korpusesse sisestamisel tuleb lähtuda autoriõigustest: vaja on teksti autori allkirjastatud nõusolekut selle kohta, et ta lubab oma tekste ja enda kohta käivat metateavet uurimistöö eesmärgil kasutada. Samuti ei saa salvestada vestlusi enne, kui kõik osalejad on andud oma nõusoleku.

Tekstide kogumine ja andmete salvestamine

- EVKK veebilehel on olemas ka ingliskeelne loa vorm, mis on tõlgitud eesti ja vene keelde ning õpilased allkirjastavad selle

Luba teksti kasutada eesti keeles

- LUBA
- Luban kasutada oma tekste teadustöö eesmärgil. Tekstid digiteerirakse ja hoiatakse Vene õppijakeele korpuse tekstiarhiivis. Neid võivad uurida ja kasutada nii Tallinna Ülikooli töötajad, nende koostööpartnerid, üliõpilased ja kraadiõppurid, kel on luba kasutada korpust.
- Tekstide koostaja kõik autoriõigused on tagatud.
- Kuupäev.....
  Allkiri.....
- TÄNAME KOOSTÖÖ EEST!
- Lisateave:
- evkk@tlu.ee

Luba tekste kasutada vene keeles

- ЗАЗРЕШЕНИЕ
- Даю свое согласие на использование моих текстов в целях научного исследования. Тексты дигитируются и хранятся в текстовом архиве корпуса русского языка учащихся эстонских школ (http://evkk.tlu.ee). Текстами могут пользоваться в исследовательских целях научные работники, студенты Таллиннского университета, а также все лица, имеющие разрешение на использование корпуса.
- Все авторские права будут соблюдены.
- Дата..... Подпись.....
- БЛАГОДАРИМ ЗА СОТРУДНИЧЕСТВО!
- Дополнительная информация:
- evkk@tlu.ee

Tekstide kogumine ja andmete salvestamine

- Tekstid on erineva pikkusega.
- Esindatud on erinevat liiki loomingulist laadi tekstid (lühikesest ümberjutustusest klassikirjandi ja esseeni).
- Tekstid on digiteeritud ja üle kontrollitud, et ei oleks mittevajalikke tühikuid, juhuslikke sisestamise apse jms.

Tekstide kogumine ja andmete salvestamine

- Iga korpuses sisalduv teksti kohta on samuti vaja fikseerida, mis teemal tekst on kirjutatud, lisada andmed tekstiliigi, abimaterjalide kasutamise kohta teksti kirjutamisel jm

Informatsioon teksti kohta eesti keeles

- Tekstiliik (essee, kirjandus, jutustus):
- Teema:
- Abimaterjalide kasutamine (sõnastikud, Internet, õpikud):
- Töö kirjutamise aeg (piiratud – klassitöö, eksamikirjandus; piiramata – kodutöö):
- Kirjutamise spontaatsus (ettevalmistatud tekst/spontaane tekst)
- Vene keele õpetaja emakeel:

Informatsioon teksti kohta vene keeles

- Тип текста (эссе, сочинение, пересказ, школьная газета и т.д.):
- Тема:
- Использование вспомогательных материалов (словари, Интернет, учебник):
- Время написание работы (ограниченное – классная работа, экзаменационное сочинение; неограниченное – домашняя работа):
- Спонтанность (текст подготовленный в классе, дома; неподготовленный)

Tekstide kogumine ja andmete salvestamine

- Kõige olulisem aspekt korpuse tekstivalikul on see, et kogutud keeleaines kajastaks sihtrühma keelekasutust adekvaatselt.
- Vaid sel juhul saab kõnelda korpuse representatiivsusest (Renouf et al., 1987; Biber et al., 1993; Sanchez, 1995; Sanchez y Cantos, 1997; EAGLES, 1996; McEnery and Wilson, 1996; Pearson, 1998; Sinclair, 1991 and 2004; etc).

Tekstide kogumine ja andmete salvestamine

- Kennedy peab oluliseks peale korpuse representatiivsuse ka tasakaalustatust ehk balanseeritust.
- Ta toetab Leechi mõtet, et korpus on representatiivne siis, kui korpusanalüüsi tulemused kajastavad üldist keelekasutust, kaasa arvatud allkeeled, žanrid ja erinevad tekstiliigid (Kennedy 1998: 62).

Tekstide kogumine ja andmete salvestamine

- Selleks tuleb kindlaks teha, missuguses proportsioonis on korpuses näiteks kirjalikke tekste ja suulist kõnet, esseesid ja ümberjutustusi või muid tekstiliiike (nt isiklik kiri).
- Reeglina on korpuses rohkem kirjalikud tekstid kui suuline kõne.
- See on seotud sellega, et kirjalikke tekste on lihtsam (vähem töömahukam) digiteerida ja keeletarkvaraga töödelda kui suulisi tekste transkribeerida ja töödelda (Kennedy 1998: 62).

Tekstide kogumine ja andmete salvestamine

- McEnery & Wilsoni arvates on lingvistid olnud rohkem huvitanud üldkeelest ja selle registrivariantidest kui üksiktekstist või ühe autori keelekasutusest.

Tekstide kogumine ja andmete salvestamine

- Representatiivsete uurimistulemuste saamiseks on oluline ka korpuse maht.
- Camino Rea Rizzo arvates ei ole veel selgelt määratud, kui palju sõnu peab olema korpuses, et see oleks representatiivne (Camino Rea Rizzo 2010: 6 (http://www.esp-world.info/Articles_27/Camino%20Rea.pdf, 22.10.2010)).
- Klassikaliselt on korpuse maht 1 miljon sõna, kuid morfoloogiliste ja süntakstiliste struktuuride, konkreetsete grammatiliste kategooriate või sõnavara uurimuseks on vaja suuremat korpust, milles oleks rohkem kui 1 miljon sõna.
- Kuna vene õppijakeele korpus on kavandatud avatud ehk monitorkorpusena, siis piirdun oma uurimuse tarvis 200 000 sõnest koosneva kirjalike tekstide koguga.

Teksti koodering (ingl *text encoding*), annoteerimine (ingl *annotation*) ja märgendamine (ingl *text markup*)

- Kolmas etapp korpuse koostamisel on kogutud tekstide ettevalmistamine uurimistööks:
- tesktid peaksid olema salvestatud ühes formaadis, mis sobib vene keele tarkvara rakendamiseks; tavaliselt txt-formaat (ingl *plain text format*)
- korpuse koostaja peab otsustama, missugust lisainformatsiooni võib vabalt korpusest kätte saada, missugune on autoriõigustega kaitstud (nt õpilase vanus, sugu, klass, teksti kirjutamise aeg ja koht jne)

- vt ka Stefanowitsch 2003: http://www-user.uni-bremen.de/~anatol/docs/corp_compilation.pdf (29.04.2010)

Teksti loodering ja märgendamine

- Korpusesse info lisamist nimetatakse korpuse märgendamiseks: mittelingvistilise info lisamist nimetatakse kitsamalt korpuse annoteerimiseks (*annotation*) ning lingvistilise info lisamist tekstidele – märgendamiseks (*tagging*).

Teksti lingvistiline märgendamine

- Teksti lingvistiline märgendamine sisaldab andmeid teksti morfoloogilise ja süntaktilise analüüsi tulemuste kohta.

Teksti lingvistiline märgendamine

- Lingvistiline märgendus võib olla tehtud käsitsi, poolautomaatselt või täisautomaatselt.
- kasutatake rahvusvahelist ühtset sümbolite keelt (nt S = substantiiv, V = verb, ADV = adverb jne).
- Aluseks võetakse keele akadeemiline kirjeldus.
- Kuigi ükski märgendusviis ei saa ennast kuulutada standardiks, kuid standardid võivad siiski hõlbustada märgendatud korpuste võrdlust jne (vt McEnery & Wilson 1996: 33 – 34).

Teksti lingvistiline märgendamine

- Vene õppijakeele korpusesse tuleb sõnaliikide märgendamine (ingl *part-of-speech tagging ehk POS tagging):*
- kus igal sõnavormil on märgend, mis näitab selle sõna kuulumist kindlasse morfoloogilisse klassi.
- Lisandub lauseliikmete märgendamine (märgendid lisatakse käsitsi)
- vealiikide märgendamine lingvistilise veaklassifikatsiooni alusel (märgendid lisatakse käsitsi).
- Vealiigimärgenduse tulemusena eraldatakse üksteisest korrektne ja normist kõrvalekalduv keel.

Teksti lingvistiline märgendamine

- Tekstid on korpuses kahel kujul:
- märgendamata ehk puhtad
- märgendatud.
- Korpusest saadakse väljavõtteid iga üksikmärgendi kaupa:
- vealiik
- sõnaliik
- lauseliige
- Vealiigi märgenduse puhul pole alust väita, et kasutatud märgendussüsteem oleks ainuõige, kuid tavaliselt on kasutatud lingvistilist veaklassifikatsiooni (aspekti valik (pöördelise vormina, infinitiivina), imperfektiivse aspekti oleviku moodustamine ja kasutamine, perfektiivse aspekti tuleviku moodustamine ja kasutamine, *винительный*-käände moodustamine ja kasutamise jne)

Annoteerimise formaadid (ingl *formats for annotation*)

- Erinevates korpusprojektides on oma märgendussüsteem. On olemas erinevaid standardeid:
  - COCOA on teatud tüüpi tekstilise informatsiooni kooderimiseks, nt autorid, kuupäevad, pealkirjad.
  - COCOA references, üks esimestest arvutiprogrammidest, kus nurksulgudes oli kahte tüüpi infot:
    - kood mis näitab teatud muutujat ja
    - tähejärjend, mis seda kirjeldab, nt <A autori nimi>

Annoteerimise formaadid

- Üldise standardina on tunnistatud TEI märgendust (*Text Encoding Initiative*), mille eesmärgiks oli välja töötada selline tesktide märgendussüsteem, mis sobiks võimalikult paljudeks eemärkideks ja oleks
- üldine, paindlik ja vajadusel laiendatav
- annaks standardse vormi, mis teeks võimalikuks teksti üleviimise ühest keskkonnast teise ja selle kasutamise teises keskkonnas
- esitaks ühtsed tekstide annoteerimis- ja märgendamispõhimõtted
- pakuks standardse vormi erinevates tekstides esinevate erinevate nähtuste märgendamiseks

(McEnery & Wilson 1996: 34 – 35; Kadri Muischneku loengud: http://www.cl.ut.ee/kursused/korp_ling05, 29.04.2010).

Annoteerimise formaadid

- TEI-d sponsoreerivad *Association of Computational Linguistics*, *Association for Literary and Linguistic Computing*, *Association for Computers and the Humanities*.
- TEI rakendamiseks kasutatakse olemasolevat dokumentide kodeerimissüsteemi SGML, mis on:
- lihtne, selge
- formaalselt piiritletud
- tunnustatud rahvusvahelise standardina
- kasutatud ka EVKK metainfo annoteerimisel
- TEI on kogum juhtnööre, kuidas seda standardit teksti kodeerimisel kasutada

(McEnery ja Wilson 1996: 35; Kadri Muischneku loengud: http://www.cl.ut.ee/kursused/korp_ling05, 29.04.2010).

Annoteerimise formaadid

- TEI formaadis koosneb iga teksti kodeering kahest osast:
- pealkirjast (sisaldab informatsiooni teksti kohta, nt autor, pealkiri, kuupäev jne; informatsioon allika kohta, nt kasutatud täpne väljaanne või kirjastaja arvutiteksti loomisel jne).
- tekstist (teksti elemendid, nt lause, lõik, peatükk jne.
- FSD (*feature system declaration*) sisaldab pealkirjas kogu asjakohast informatsiooni TEI kodeeringus. TEI-s tekstid põhinevad DTD-l (*document type description*), mis ütleb arvutikasutajale või arvutirprogrammile, missugustest osadest tekst koosneb ning kuidas need osad on üksteistega seotud. Nn TEI-LITE on standardiseeritud kogum kõige tavalisematest või tähtsamatest TEI märgistest. Seal on kasutatud ainult osa juhtnööridest

(McEnery & Wilson 1996: 35 – 37).

Kokkuvõtteks

- Korpuse koostamine hõlmab mitmeid valikuid, mis on omakorda seotud korpuse olemuse ning tüübiga.
- Vene õppijakeele korpus kuulub spetsiaalsete korpuste alla, on monitorkorpus, sisaldab gümnaasiumiõpilaste vene keelt esimese ja teise keelena.
- Tekstid on digiteeritud doc- ja txt-formaadis, mis on üldtunnustatud standard. Tekstid peaks olema märgendatud ning annoteeritud.

Kirjandus

- **Eslon, P & Metslang, H. 2007.** Õppijakeele ja eesti vahekeele korpus. – Eesti Rakenduslingvistika Ühingu Aastaraamat (III). Tallinn: EKS, 99 – 116.
- (http://evkk.tlu.ee/wwwdata/pdfs/oppijakeel.pdf, 01.05.2010).
- **Eslon, P. 2007.** Õppijakeelekorpused ja keeleõpe. – Tallinna Ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine / Toim. P. Eslon. Tallinna Ülikooli eesti filoloogia osakonna toimetised 9. Tallinn: Tallinna Ülikooli Kirjastus, 87 – 120.
- http://evkk.tlu.ee/kogumik2007/korpusekogumik_pille.pdf (01.05.2010).
- **Kennedy, G. 1998.** An introduction to corpus linguistics. New York: Longman.
- **Kitsnik, M. 2007.** Õppijakeele uurimine ja arendamine – põnev väljakutse. – Emakeel ja teised keeled. V. Tartu : Tartu Ülikooli Kirjastus, 41 – 55.
- http://evkk.tlu.ee/pdfs/Kitsnik_artikkel151106.pdf (29.04.2010).
- **Копотев, М.В. 2003.** Корпусная лингвистика в Финляндии (обзор ресурсов). – Научно-техническая информацияю Серия 2. №6, 37 – 43.
- http://www.helsinki.fi/~kopotev/finnish_corpora.pdf (28.04.2010).
- **McEnery, T. & Wilson, A. 1996.** *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- **Sõrmus, K.** 2008. Emakeeleõppija koprus. Statistiline analüüs ja veamärgendussüsteem. Magistritöö. Juh. Kadri Muischnek (PhD), Maia Rõigas (dots). Tartu.

Internetiallikad:

- **Camino Rea Rizzo.** 2010. Getting on with corpus compilation: from theory to practice. vt http://www.esp-world.info/Articles_27/Camino%20Rea.pdf (22.10.2010).
- **Kadri Muischnek,** Korpuslingvistika, loengud
- vt http://www.cl.ut.ee/kursused/korpuslingvistika (29.04.2010).
- **Stefanowitsch, A. 2003.** *Corpus Compilation*. Corpus linguistics. vt http://www-user.uni-bremen.de/~anatol/docs/corp_compilation.pdf (29.04.2010).

Edasi lugemiseks:

- **Barlow, Michael.** 2005. Computer-based analyses of learner language. – Analysing Learner language. Barkhuizen, Ellis. Oxford: Oxford University Press.
- **Bowker, Lynne, Jennifer Pearson**. 2002. Working with Specialized Language. A practical Guide to using Corpora. Bury ST Edmuds, Suffolc, Great Britain: ST Edmundbury Press.
- **Granger, Sylviane**. 2003. A promising Sunergy. – Error-tagged Learner Corpora and CALL: Calcio 20 (3).
- **Granger, Sylviane.** 2006. Center for English Corpus linguistics. http: //fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/research%20learner%20corpira.html
- **Granger, Sylviane.** 2007. A bird-eye view of learner corpus research. – Corpus linguistics. Critical Concepts in Linguistics vol VI. Ed by Wolfgang Teubert & Ramesh Krishnamurthy. Routledge.
- **Sinclair, John**. 1995. Corpus, Concordance, Collocation. Oxford: Oxford University Press.
- **Sinclair, John**. 2004. How to use Corpora in language Teaching. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Tänan tähelepanu eest!
Thanks for your attention!
Kiitos huomiosta!
Благодарю за внимание!